A.6 Approved Work Plans and Deliverables

For year 2 and year 3, the work plans, including milestones, are subject to modification by mutual agreement of the Library and the awardee. These modifications will be based on what is learned during year 1.

WORK PLAN FOR CONTENT IDENTIFICATION AND SELECTION

General Model for Content Identification and Selection

Following is a series of activities pertaining to Content Identification. It is, to varying degrees, reiterative throughout the project, as it is expected that partners will emerge over time, and to some extent due to the engagement of early adaptors or recruitment on the part of Advisory Group members. It is also expected that parameters will evolve as the result of experience and the specific strengths of future repository partners.

- Refine parameters for selection
    - Must be geospatial
    - Data pertaining to North America takes precedence
    - Must be in digital form
    - Multiple data types allowed
- Develop procedures for obtaining data and related information from content provider
    - Develop a questionnaire for determining such information as:
        - What is the topic of the digital data to be preserved?
        - What is its importance? That is, why is it important to the collective holdings of the United States? Why is it of long-term interest to the Library of Congress? How is it currently used? If it is not currently used, what is its potential usage?
        - Are the data available to the public?
        - How is the data at risk? What needs to be done to protect it?
        - What data types are included?
        - What is the size of the files/ collection?
        - What is the quality of metadata?
        - Other issues
- Identify initial group of content providers to solicit (see lists below)
- Make contact with each proposed content provider (if not already engaged)
    - If identified content provider is willing to participate, adjust work plan accordingly
- Follow-up at least twice if there is no response to initial solicitation
- Send out blanket requests for participation to appropriate listservs
    - Identify appropriate listservs
    - Write requests
    - Submit requests

- Make requests for participation at appropriate conferences, meetings, etc.
    - Identify appropriate conferences, meetings, etc.
    - Identify person to make announcement
- Engage advisory group
    - Itemize responsibilities of advisory group members
    - Compose letter describing responsibilities
    - Finish recruiting advisory group members
    - Ask for advisory group's input in identifying more content that should be included in the project
    - Enlist their aid in making contacts, as appropriate
    - Enlist their aid in making selection decisions about whether or not to include content which was not solicited directly
    - Advisory group will help decide policy issues, e.g. will all content be accepted
        - If not, what criteria are applied to make inclusion/ rejection decisions
- Importance of data?
- How at risk is the data?
- Quality of data?

- Construct tables of information about prospective collections, including:
    - Subject
    - Display characteristics
    - Use characteristics
    - Size of collection

- Document action plan for handling content provider, based on information gathered from the questionnaire filled out by each content provider (with technical assistance from staff at UCSB or Stanford as needed) with such detail as:
    - To which repository the data will be transferred
    - Procedures for transferring data from the content providers to the repository
        - Specific to the requirements of each content provider, data type(s), and available metadata

Some of the target content is fixed and complete, e.g., scanned or satellite images, and we will capture these through orderly, inventory-controlled data exchange with the producer. Other target content is periodic in nature, but relatively stable, e.g., GIS data released by state agencies, which make it possible to harvest on schedule (e.g., through LOCKSS).

Satellite images will be either from completed missions or from missions in progress; in the former case, the images may be archived as a “closed” archive, and in the latter cases the images will be added to in a regularly scheduled upload of data and existing metadata. GIS layers generated by a county government by definition are continuously updated; a

major decision point will be how much change, and/or what kind of change, requires a new snapshot to be added to the archive.

Still other target content is additive, but irregular, e.g., newly scanned images produced by partners, which require ongoing communication and collaboration. Finally, some target content is neither fixed nor final, e.g., municipal GIS data where currency is of the highest importance to the owner/producer. In such cases, the project will need to discuss and devise standards and practices for periodic snapshots, such that we collect the most current version on a scheduled basis without overwriting older versions that may be of future research, policy, or even forensic interest.

In terms of topical content parameters, the program will acquire geographical, geological, environmental, resource management, and related information pertaining mainly to North America (primarily the United States), both past and present. Some of this will have been professionally published either in analog or digital form; some will not. Much of it will be government-produced (at local, state or federal level); some will be university-based; some will be society-produced; some of it will be commercial. The project will develop and negotiate a small, but flexible, set of rights agreements that will assure cooperation with the producer/owners, whatever their nature or business models.

Stanford and UCSB will divide content identification and selection between them, with UCSB emphasizing born-digital material (such as satellite images - an area of specialty for UCSB - and the GIS layers of Santa Barbara County, California, county agencies) and interaction with, and mining of, two classic Websites (Oddens' Bookmarks, http://oddens.geog.uu.nl/index.html; Map History, http://www.maphistory.info/) that are recognized internationally as sources of URLs for digital geospatial data, and that frequently and regularly seek out new sites on the Web.

Topical content will in the case of satellite images be base data in the main (e.g., remote-sensing images in original format); GIS layers may be of any theme that county workers need to understand the needs of the county; and data mined from the two URL Websites will generally be maps, with a focus on maps documenting important social and political developments and foreign affairs, in the main from URLs obtained via Oddens' Bookmarks and Map History.

In terms of technical content parameters, the intention is to acquire and retain content in many digital forms and formats. However, with the certainty that some formats will become obsolete and thus unreadable faster than other formats, there will be a continuous dialog among the media preservationists, content owners, repository managers and technologists about migration or translation of data. In brief, selection of content will not be based on the formats acquired, and thus retention of that content will have to accommodate a large variety of complex formats.

While the emphasis is on collecting data in non-proprietary formats, a major challenge will be determining how to deal with proprietary formats and therefore with proprietary software.

Confirmed Sources:

- Satellite images as currently within the scope of the Alexandria Digital Library.
- GIS layers of Santa Barbara County, California, county agencies
- Oddens' Bookmarks, http://oddens.geog.uu.nl/index.html

- Map History, http://www.maphistory.info/
- 8,800 historical maps, mainly of North America, from the David Rumsey Collection, Cartography Associates, including full bibliographical information (expected to grow during project by about a thousand maps per annum); currently ca. 2 terabytes, http://www.davidrumsey.com
- 1,000 historical maps of the Early Washington Maps Project, produced by Washington State University, the University of Washington, and Tacoma Public Library, http://www.wsulibs.wsu.edu/holland/masc/xmaps.html
- 1,500+ historical maps 16th-20th centuries from the Hargrett Rare Book Lib of the U of Georgia, with emphasis on Revolutionary War era Georgia and colonial America. http://www.libs.uga.edu/darchive/hargrett/maps/maps.html
- 1,500 historical map and field notebook images from the Stanford Geology Survey, http://gill.stanford.edu/depts/branner/SGS_home.html
- GIS and other files from the Indiana State Geological Survey
- All GIS and selected other files available to the public from the California Resources Agency, http://www.gis.ca.gov
- 72 MODIS composite files (comprising one terabyte of data) held by the University of Maryland Global Land Cover Facility: http://glcf.umiacs.umd.edu/index.shtml
- 350 USGS topological maps of the San Francisco Bay Region (1897-1997) scanned by the University of California, Berkeley, Earth Sciences and Map Library http://sunsite.berkeley.edu/histopo/
- 13,500 photographs of American gravestones from the Forbes and Farber collections of the American Antiquarian Society, including transcriptions ; ca. one-half terabyte
- GIS databases from the extensive collections and working sets of the Stanford University School of Earth and Environmental Sciences
- Maps, databases, and other results of geographical research conducted at Stanford
- Geospatial data held by the Stanford Earth Sciences Library, including, but not limited to: California Landsat imagery, DOQQs (digital orthophoto quarter quadrangles) covering California, SPOT Statewide California remote-sensing images (copyrighted data), plus GIS and aerial photography. Total identified to date, ca. 8,500 files or one-half terabyte.

Sources Currently in Negotiation:

- Maps, files, databases and journals from scholarly societies, e.g.:
    - American Geological Institute
    - American Geophysical Union
    - Association of American Geography
    - Geological Society of America
- Campus maps, GIS data from the Stanford University Architect / Planning Office
- A sample of historical map and related data files from the G&M department of the Library of Congress
- Historical maps, esp. of the Mid-Atlantic States, from New York Public Library

- Maps, files, and databases produced or aggregated and curated by the United States Geological Survey, National Geologic Map Database Project
- ESRI: GIS and related data files aggregated by its subsidiary GeographyNetwork.com
- Nevada Bureau of Mines

Prospective Sources:

- NASA
- USGS
- Maps.com
- Maps and other files (primarily GIS data) compiled by state Geological Surveys. Ideally, we would acquire maps and GIS data from all state Surveys, but we are particularly hopeful of working with those perceived to be in dire budget situations (i.e., most at risk of being defunded), including:
    - California
    - Washington
    - New York
    - Connecticut
- Local government GIS and other cartographic information

Content Identification, Selection, and Acquisition Milestones

- First year
    - First quarter
        - First 7 weeks:
            - advertise, interview, and hire project staff
            - develop and implement procedures for identifying and selecting data, and for acquiring from content providers data and metadata; discuss with Oddens' Bookmarks, Map History, NASA, ESRI, and other primary sources of digital geospatial data to determine general schedule and work flow for obtaining site URLs or data
            - write form emails/letters to send to owners of digital geospatial data re possible archiving of data
            - Recruit and engage Advisory Group
            - Get Webpage/Webpage architecture for grant set up; write procedures for additions by grant personnel
        - Remaining 6 weeks:
            - first week: orientation for new staff
            - next 5 weeks: train new staff and document processes
    - Second quarter: programming staff from UCSB travel to Stanford to assist in loading ADL-Operational on Stanford servers; pilot tests of obtaining content from content providers (e.g., California state and selected county government agencies; URLs obtained from Oddens' Bookmarks and Map History; loading of data and metadata into UCSB ADL begins; begin

writing procedures for content identification, selection, and acquisition (including loading procedures)
    - Third quarter: pilot tests of loading data and metadata into ADL-Operational at Stanford (programmers from UCSB to go to Stanford to assist); re-write and improve procedures; preparation of demonstration for Advisory Board and for workshop of grant personnel
    - Fourth quarter:
        - by end of fourth quarter, aim to have two terabytes of archived digital geospatial data
        - hold 1 workshop, combining grant personnel and libraries interested in building archives of digital geospatial data and participating in a national network of such archives; to evaluate work done and to do in coming four quarters (and more generally in second year of grant)
        - following workshop, hold meeting of Advisory Board
- Second year
    - First quarter: from lessons learned during the first year, determine content identification, selection, and acquisition priorities for each quarter of second year of grant; identify and select content as per parameters; obtain data and metadata
    - Second quarter: fine-tune identification and selection parameters; set loading priorities; evaluate loading procedures for data and metadata
    - Third quarter: content identification, selection, and acquisition continues; start writing best-practices documents, aiming toward draft documents for workshops held at end of fourth quarter; best-practices document to include how other libraries may load ADL-Operational and build their own geospatial-data archives
    - Fourth quarter
        - by end of fourth quarter, aim to have four terabytes (total) of archived digital geospatial data
        - hold 2 workshops on archiving digital geospatial data, one for grant personnel and one for libraries interested in building digital geospatial-data archives and participating in a national network of archives of digital geospatial data, with the draft best-practices document as chief topic of work for both workshops
        - hold meeting of Advisory Board

- Third year
    - First quarter: re-write draft best-practices document into a solid Version 1.0 document; continue content identification, selection, and acquisitions
    - Second quarter: issue Version 1.0 best-practices document for comment
    - Third quarter
        - hold one workshop on archiving digital geospatial data, for libraries interested in building archives of digital geospatial data and in participating in a network of archives of digital geospatial data; Version 1.0 document as chief topic of workshop

        - hold meeting of Advisory Board
    - Fourth quarter
        - by end of fourth quarter, aim to have six terabytes (total) of archived digital geospatial data
        - by end of fourth quarter, have "final" version of best-practices document - to include how to become a member of a national network of archives of digital geospatial data - available via the grant's Webpage
        - get all final reports written

WORK PLAN FOR CONTENT ACQUISITION

General model for content acquisition

Each of the Partners will have its own technologies, procedures, and challenges for content acquisition. As indicated in the general work plan for content identification and selection, acquisition procedures will be driven very much by the nature of the source and format of the content and its associated metadata. One of the project's objectives is to automate the metadata extraction as much as possible.It is likely that the content acquired by UCSB will lend itself more readily to such automation than that targeted by Stanford. For milestones, see milestones for Content Identification and Selection.

UCSB Content Acquisition Work Plan

UCSB Alexandria Digital Library (ADL):

ADL has an operational Catalog (ca. 2 million metadata records) and Gazetteer (ca. 5 million place-name records) in public service. Since 1994 ADL has ingested more than eight terabytes of content and over six million metadata records. This experience has born a robust processing production line and many automated processes, which will be used for this project.

ADL has two work flows; one for data and the other for metadata. The data workflow consist of: data object analysis; extracting metadata from headers if available; processing the raw data objects into Internet deliverable objects; constructing scripts for automated generation of thumb and browse images; establishing of geographic coordinates; then moving object to collection directory trees. The metadata workflow consists of: obtaining and analyzing available metadata; generating missing search field content; converting database records into XML reports; building search tables; connecting metadata records with data objects and derivatives; and quality proofing Internet searching and object delivery.

Following are the current summary instructions for installing a geospatial collection, which will serve as a model for the ingestion of content sets under the proposed archive:

1. Create collection-level metadata by completing specialized forms in a web page.
2. Create or map item-level metadata to the ADL standard views.
3. Create configuration files.

4. Locate your data objects, thumbs, and browse images.
5. Add collection configuration, and test
6. Commit configuration source code repository.
7. Add collection to production services by pulling configuration from source code repository.

Architecture of the ADL technical infrastructure

The overall architecture of the Alexandria Digital Library comprises three components: client(s), middleware and server. The ADL middleware server mediates between digital library clients and digital library collections. This architecture will support the content acquisition for the project.

Middleware Server

The ADL middleware server is a distributed, peer-to-peer software component that provides mediated programmatic access to digital library collections. To clients it presents standard library services in the areas of metadata search, retrieval, and ranking; access control; and collection management and organization. To collections it presents a standard framework in which heterogeneous, collection- and/or item-specific metadata can be mapped and returned. The middleware server works cooperatively with a centralized collection discovery server and, in peer-to-peer fashion, with other middleware servers. The server is written in Java and Python and can be run as a web application inside a servlet container, as an RMI (Remote Method Invocation). server, or both. Distributed with the server is the "Bucket99 driver," a configurable component that allows relational databases to be viewed as collections.

The middleware's query language is a generic, XML-encoded language that supports boolean combinations of typed constraints against abstract metadata fields, or "search buckets." The syntax and semantics of the language are defined by a DTD. ADL has defined a set of nine standard search buckets. These buckets are defined conventionally, not architecturally. Nevertheless, clients can assume that all collections support all of the following buckets:

Geographic locations	the item's spatial footprint, i.e., an approximation of the subset of the Earth's surface to which the item is relevant, expressed as any of several types of geometric regions defined in WGS84 latitude/longitude coordinates.
Dates	the item's temporal footprint, i.e., the range of calendar dates to which the item is relevant.
Types	terms drawn from the ADL Object Type Thesaurus identifying the meaning or content of the item.
Formats	terms drawn from the ADL Object Format Thesaurus identifying the form or representation of the item.
Titles	the item's title. This bucket is a subset of the "Subject-related text" bucket.

Originators	names of entities related to the origination of the item (authors, publishers, distributors, etc.).
Assigned terms	subject-related terms from controlled vocabularies. This bucket is a subset of the "Subject-related text" bucket.
Subject-related text	text indicative of the subject of the item, not necessarily from controlled vocabularies. This is a superset of the "Titles" and "Assigned terms" buckets.
Identifiers	item names and codes that serve as unique identifiers

Middleware Server Client Interfaces

There are three client interfaces to the middleware: Java (which supports direct access to Java objects and methods), HTTP, and RMI. The three interfaces are each partitioned into ten independent, stateless services. The principal services include:

configuration()	returns middleware configuration properties. The property of most interest to clients is the list of available collections.
collection(collection-name)	returns collection-level metadata for a collection.
query(query)	asynchronously queries one or more collections. query is a query expressed in the middleware's query language. Results are stored in a "result set" on the server. query-id identifies both the running query and the result set.
results(query-id)	returns a reference to a result set. A result set is a server-side object that holds the results of a query in the form of a single merged, ranked list of collection items, where each collection item is represented by a triplet of the item's three standard ADL metadata views. Result sets carry sundry other properties such as elapsed time, total number of results processed, etc.
metadata(view, collection-name, item-identifier)	returns a view of the metadata for an item within a collection.

Additional services are under development that will support collection-level discovery (i.e., discovery of relevant collections) and collection building and management.
All metadata returned by the middleware is encoded in XML and adheres to publicly-defined, external DTDs which contain human-readable documentation, and define both syntactic form and semantic content.
The collection-level metadata for a collection contains a number of items of interest to

clients: the buckets supported by the collection; the thesauri used to categorize the collection; the number of items in the collection both overall and broken down by type and format; the spatiotemporal distribution of the items; and so on. Item-level metadata is packaged into "views," with the views differing in size, content, encoding style, and intended usage. Collections and even individual items within collections may define and return arbitrary views. ADL defines three standard views which all collections and items must support:

Bucket: describes mappings from the item's native metadata to high-level, searchable buckets.
Browse: describes the browse-size representations of the item that are available.
Access: provides the information necessary to download or otherwise access the item's content.

Middleware Server Collection Interface

Conceptually, an ADL collection is 1) a set of items, each of which has item-level metadata and an identifier that is unique within the collection; together with 2) collection-level metadata. Functionally speaking, an ADL collection is a set of services that return collection-level metadata; that return various standard views of item-level metadata; and that, most significantly, query the collection and return items matching one or more constraints placed against item-level metadata. Although not required logically, it is expected that collections respond to requests quickly, consistently, and reliably, and that identifiers have relatively long-term persistence.
The interface to a collection is implemented by a driver, or more precisely, by three drivers, which respectively implement the middleware's collection, metadata, and query services for that collection. The middleware dynamically and independently loads drivers on demand, and may unload a driver at any time.
ADL has specified a content standard and XML encoding for collection-level metadata. Collections are generally free to use any item-level metadata, the only requirement being that item-level metadata must be mapped to the ADL search buckets and to the standard ADL metadata views. Collections may also return other views of item-level metadata not defined by ADL.

Stanford Digital Repository: General Description

Content collected by Stanford through and for the project will be ingested into the Stanford Digital Repository (SDR). The Digital Repository houses collections of digital objects as well as metadata about the individual objects and collections of objects. Digital objects include both objects that originate in digital form and ones that are digitized versions of analog objects.
The Digital Repository is composed of several different functional layers. Each layer is implemented using third-party software. This approach reduces the staffing necessary to implement or enhance the repository, as well as providing maximum flexibility in using new tools and technology as they arise.

The layers of the Repository are:

- A discovery layer, the web page where the researcher searches and browses the digital collections;

- A metadata manager, which keeps track of the information about digital objects and collections of objects in the repository;
- An object manager, which keeps track of the items in the collection;
- An object repository, where the items in each collection are kept; and
- Management tools, which enable library staff to add digital objects to the repository and add/modify metadata about each object or collection.

The Discovery Layer

Stanford will use Lucene, the search tool used by DSpace, as the core of the discovery layer. Because Lucene is a set of search API's rather than a distinct search engine, we have the ability to tailor the search criteria for each type of object type present in the repository. Thus a search for PDF's can be based on a phrase found within the text of the PDF, while a search for maps can be based on the resolution of the map's resolution.

The Metadata Manager

All technical and structural metadata are maintained in METS records. In addition, pointers to specialized descriptive metadata - such as EAD, MARC, Dublin Core, or their successor formats - can be wrapped in the METS record. Because METS is an XML construct, we have chosen to store the METS records in an XML database. Tamino, by Software AG, is the software product chosen for this task. For more on metadata, please refer to the section on metadata in the workplan for Content Acquisition, above.

The Object Manager

An object represented by a METS record may in fact consist of several file objects. As an example, consider the case of a text document. The document will be fully described by a METS record, but resides in the repository as a number of individual files: A TIF image file for each page in the document, plus a PDF file as the delivery surrogate which most patrons will use, plus a number of text files representing the XML-marked-up pages of the document.
The object manager keeps track of the interrelationship of each of these files. This allows the METS record for the object to hold one unique ID representing the document, while letting the object manager deal with the more complex interrelationships. This also provides a convenient way to offer only the set of files appropriate for the patron requesting the object.

At the time that a request is made, the ID of the requester is compared to the access rights for the particular object. Only the files which a user has the right to access are made available. Thus while most requests will result in display of the PDF file, scholars may also have access to the original TIF (or other) files.

Stanford has chosen TEAMS, a Digital Asset Management system, as the object management software. This system is functioning according to expectations. However, it is assumed that it will eventually be replaced by some other system and accordingly, the implementation of the Repository is not to be dependent on specific features or standards inherent in the TEAMS system.

The Object Repository

Ultimately the objects in a collection must reside on-disk somewhere. That is the job of the Object Repository. Stanford is using an EMC Centera server to store the delivery surrogates.
In cases where the delivery surrogate is not also the source object, the source objects are kept in near-term storage such as slower disk or a tape library.

Stanford Content Acquisition - Metadata Focus

The capture and ingestion procedures for the SDR extract as much of the metadata related to structural / technical identification as possible from the digital objects themselves using software tools designed and/or adapted for that purpose at both the capture and the ingestion point. We have worked with a number of software vendors as well as service providers to adapt their products or services to as much automated metadata capture as possible. We also capture key identifying (descriptive) metadata that will allow us to provide at least a minimal level of retrieval along with a persistent location within the repository. Along with at least the minimal descriptive, technical and structural information we try to assess the level of preservation risk associated with a given unit or class of digital object following the principles described by William G. LeFurgy in his article, "Levels of Service for Digital Repositories"
http://www.dlib.org/dlib/may02/lefurgy/05lefurgy.html .

As with other formats, staff will decide on minimal levels and/or required metadata elements critical to providing the appropriate level of service for long-term access to the different formats in the geospatial collection. Formats are expected to include spatial imagery (Landsat, MODIS, DOQQ), data in a GIS format, and conventional images (TIF, .jpg, Mr. SID) and metadata created using Dublin Core, MARC format, or the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) and the ISO 19115, Geographic Information - Metadata.

The other area of critical importance to capture at an early phase of the ingestion process is related to rights, specifically who are the rights holders for the content, and what are the permissions and constraints related to the short and long term usage of the geospatial content. Preferably, this information would be collected at the time the service agreements are signed with content owners, and then it would be a relatively simple procedure to incorporate the information into the Rights schema that the SDR is using, and associate that with the other metadata within the METS records for each digital object.

The Stanford Metadata Librarian and the GIS & Map Librarian and the Assistant Head of UCSB MIL will work closely together to develop metadata standards for geospatial content, an area in which the staff of the ADL are recognized experts. This expertise will be critical as Stanford begins archiving contributed content. The three librarians will do most of the analysis of collection level and format dependent metadata that will be needed for ingestion into the SDR prior to invoking of the ingestion procedures. Of critical importance will be getting sample content and metadata from each of the subsets of the geospatial collection coming to Stanford in order to allow the most efficient automated processing of each sub-collection by processing staff.

Metadata Strategies for Geospatial Data

The mission of the SDR is to provide long term retention, storage, and preservation of the digital materials in the repository. To that end, SDR staff have either required or recommended various types and extent of metadata including not only the traditional discovery or bibliographic metadata, but also technical, rights, structural, and preservation metadata that is necessary to provide technical sustainability of digital collections over the long run. To assure content providers that the SDR is following best practices and guidelines for digital repositories, various staff have been participating in and/or monitoring discussions related to the establishment of criteria for a "trusted digital repository" as described in the RLG/OCLC report entitled: "Trusted Digital Repositories: Attributes and Responsibilities" (See http://www.rlg.org/longterm/repositories.pdf.) SUL's intention is to conform to the expectations expressed in that document as finalized, and become certified as a trusted digital repository.

Geospatial Metadata: Descriptive, Technical and Rights

Metadata standards currently in use or being considered for the SDR include MARC, Dublin Core qualified and simple, and MODS for descriptive metadata, the NISO Z39.87 standard for technical metadata for still images, and the METS extension schema for technical metadata for text. Because of the variety of formats that would be included in the geospatial collection, SDR Metadata staff, in conjunction with the GIS & Map Librarian and the Assistant Head of MIL, would need to assess whether these standards should be applied to the geospatial content, or whether other standards would be best used for the spatial imagery (Landsat, MODIS, and DOQQ) and data in a GIS format. We will consider a variety of metadata formats including UCSB's ADL standard, the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) and the ISO 19115, Geographic information - Metadata. (See http://www.fgdc.gov/metadata/meta_stand.html).
Of particular interest is the work done at Cornell University in this area as described in a paper from the Dublin Core Conference 2003 (http://dc2003.ischool.washington.edu/).:

Westbrooks, Elaine L. "Efficient Distribution and Synchronization of Heterogeneous Metadata for Digital Library Management and Geospatial Information Repositories" http://www.siderean.com/dc2003/204_Paper78.pdf

It is hoped the Partners' Technical Services staff and maps librarians will engage substantively with the staff of the Cornell University Geospatial Information Repository in refining standards for metadata to be applied to the project.

Geospatial Metadata: Structural

In terms of structural metadata, the SDR is using the emerging Metadata Encoding and Transmission Standard (METS) schema (http://www.loc.gov/standards/mets/)to describe and manage the components and associated files of the digital objects within the repository. Maintaining the structural relationships among associated files is of critical importance to geospatial and imagery data, so our experience with METS, and the fact that the SULAIR Metadata Librarian sits on the METS Editorial Board should greatly facilitate the expansion into the use of METS for geographic / geospatial data.

Geospatial Metadata: Preservation

Another very important type of metadata that will be necessary to determine for the geographic data will be preservation metadata. Currently, there is a great deal of discussion in the field about important digital preservation issues such as what are the rules and guidelines for a digital repository, what constitutes preservation metadata, what are core preservation metadata elements, and how are / can they be tested in actual implementation scenarios. A number of working groups have been formed in digital communities around the world that are being monitored by SDR Metadata and Preservation staff. SULAIR's Metadata Librarian is an active participant on the OCLC / RLG Preservation Metadata Working Group (PREmis) whose focus is defining workable implementation strategies using the first Working Group's "Metadata Framework for Supporting Preservation of Digital Objects" (http://www.oclc.org/research/projects/pmwg/default.htm). The results of this Working Group's efforts will inform SULAIR decisions about which core preservation metadata elements are advisable for geographic data from a theoretical perspective. To jumpstart the establishment of a practicable strategy, SULAIR would have the opportunity to test the efficacy of the core metadata elements to facilitate preservation should we be able to offer some of the content associated with the geospatial collection as one of the pilot preservation implementation programs for the PREmis Working Group.

Content Acquisition Milestones

- First year
    - First quarter
        - First 7 weeks:
            - advertise, interview, and hire project staff
            - develop and implement procedures for identifying and selecting data, and for acquiring from content providers data and metadata; discuss with Oddens' Bookmarks, Map History, NASA, ESRI, and other primary sources of digital geospatial data to determine general schedule and work flow for obtaining either URLs of sites or data
        - write form emails/letters to send to owners of digital geospatial data re possible archiving of data
        - Recruit and engage Advisory Group
        - Get Webpage/Webpage architecture for grant set up and live on the Web; write procedures for additions by grant personnel
            - Remaining 6 weeks:
        - first week: orientation for new staff
        - next 5 weeks: train new staff and document processes
    - Second quarter: programming staff from UCSB travel to Stanford to assist in loading ADL-Operational on Stanford servers; pilot tests of obtaining content from content providers (e.g., California state and selected county government agencies; URLs obtained from Oddens' Bookmarks and Map History; loading of data and metadata into UCSB ADL begins; begin

writing procedures for content identification, selection, and acquisition (including loading procedures)
    - Third quarter: pilot tests of loading data and metadata into ADL-Operational at Stanford (programmers from UCSB to go to Stanford to assist); re-write and improve procedures; preparation of demonstration for Advisory Board and for workshop of grant personnel
    - Fourth quarter:
        - by end of fourth quarter, aim to have two terabytes of archived digital geospatial data
        - hold 1 workshop of grant personnel and libraries interested in building archives of digital geospatial data and participating in a national network of such archives; to evaluate work done and to do in coming four quarters (and more generally in third year of grant)
        - following workshop, hold meeting of Advisory Board
- Second year
    - First quarter: from lessons learned during the first year, determine content identification, selection, and acquisition priorities for each quarter of second year of grant; identify and select content as per parameters; obtain data and metadata
    - Second quarter: fine-tune identification and selection parameters; set loading priorities; evaluate loading procedures for data and metadata
    - Third quarter: content identification, selection, and acquisition continues; start writing best-practices documents, aiming toward draft documents for workshops held at end of fourth quarter; best-practices document to include how other libraries may load ADL-Operational and build their own geospatial-data archives
    - Fourth quarter
        - by end of fourth quarter, aim to have four terabytes (total) of archived digital geospatial data
        - hold 2 workshops on archiving digital geospatial data, one for grant personnel and one for libraries interested in building digital geospatial-data archives and participating in a national network of archives of digital geospatial data, with the draft best-practices document as chief topic of work for both workshops
        - hold meeting of Advisory Board

- Third year
    - First quarter: re-write draft best-practices document into a solid Version 1.0 document; continue content identification, selection, and acquisitions
    - Second quarter: issue Version 1.0 best-practices document for comment
    - Third quarter
        - hold one workshop on archiving digital geospatial data, for libraries interested in building archives of digital geospatial data and in participating in a network of archives of digital geospatial data; Version 1.0 document as chief topic of workshop
        - hold meeting of Advisory Board

- Fourth quarter
  - by end of fourth quarter, aim to have six terabytes (total) of archived digital geospatial data
  - by end of fourth quarter, have "final" version of best-practices document - to include how to become a member of a national network of archives of digital geospatial data - available via the grant's Webpage
  - get all final reports written

WORK PLAN FOR PARTNERSHIP BUILDING

General Model for Partnership Building

The Partners' goal is to design and establish a framework for a distributed digital repository for geospatial data for the general good, on behalf of many communities, indefinitely into the future. We intend that the Repository fill an active and cooperative role in the nascent national program of digital repositories as envisioned by the Library of Congress through NDIIPP. We will work together closely - to share practices and experience, to foster standards, to explore effective interoperation, and to recruit additional partners.

Clear and open written agreements will be crucial to the content provider partners. Broadly, there is little trusted third-party digital archiving tradition or established practice (ignoring in this context corporate data warehousing or remote storage of business records); this project - and others within NDIIPP - will be forging the relationships and supporting instruments for such a practice. Having worked with about 150 publishers participating with the HighWire Press and/or LOCKSS, Stanford has considerable experience in the respectful treatment of publisher rights, interests and concerns, on the basis of which no serious difficulties are anticipated in establishing the Partners as trusted third-party repositories for digital geospatial content.

Content provider partners are in different states of readiness to act, whether for technical or deliberative reasons. The plan below suggests the general sequence of events.

- Solicit a pool of early adaptors from several sectors (September 2003 - June 2004)
  - Universities
  - Professional / Scholarly societies
  - Federal agencies, esp. USGS, NASA, NIMA
  - State Geological Surveys
  - Municipal and county planning agencies
  - Other agencies and data holders
- Adapt, negotiate and execute several standard versions of a repository agreement (April - June 2004)
  - Retained university IP counsel will coordinate this effort
- Negotiate transfers of sample data from partners (April 2004 - ongoing)
  - Content objects as well as metadata (in whatever form)
  - Clarify format and linking aspects of partner data

- Develop production parameters and schedule for each partner - batches and/or ongoing capture (April 2004)
- Perform detailed tallies, analysis and quality control of provided content (Ongoing once data have begun to arrive in production mode)
    - Renegotiate with provider as necessary
    - Work with provider to clarify or solve content-specific issues
- Report to provider
    - on completion of discrete bodies of information
    - on agreed schedule for ongoing ingestion

The work plan for provider partners may be affected by discussions with the Library of Congress, particularly with regard to standards and expectations.

Growth Plan for Addition of Repository Peer Partners

The goal for this aspect of the project will be to recruit to the partnership several well-established digital repositories with geospatial content in the United States. These partnerships may include ingesting imagery into the Repository, building technical partnerships for experimentation with interoperability, establishing best practices for metadata standards, or creating models for migration of complex data structures, such as banded imagery or GIS data files. Potential partners include:

- The National Geospatial Data Clearinghouse (http://130.11.52.178/gateways.html): a collection of over 250 spatial data servers. Downloadable content will be harvested ad archived from participating sites.
- The Geography Network (http://www.geographynetwork.com/data/index.html): ESRI's data clearinghouse and map serving Web site. This may need to be a set of replicated database, or their content.
- CIESIN (http://www.ciesin.org/): The Center for International Earth Science Information Network at Columbia University.
- The USGS EROS Data Center (http://edcwww.cr.usgs.gov/products/satellite.html): Includes aerial photography, satellite imagery, elevation and land cover data, and maps.
- CUGIR (http://cugir.mannlib.cornell.edu/): Cornell University Geospatial Information Repository, an FGDC Clearinghouse Node for New York State.

Partnership Building Milestones

- First year
    - First quarter
        - First 7 weeks:
            - advertise, interview, and hire project staff
            - develop and implement first draft of procedures for identifying and working with potential partners
            - write form emails/letters to send to potential partners (owners of digital geospatial data) re possible archiving of data

            - Get Webpage/Webpage architecture for grant - with sections for potential and committed partners - set up and live on the Web
            - Begin planning for first partners meeting
        - Remaining 6 weeks:
            - first week: orientation for new staff
            - next 5 weeks: train new staff and document processes; continue planning for first partners meeting
    - Second quarter: pilot tests of working with potential partners (content providers such as California state and selected county government agencies); write second draft of procedures for working with potential and committed partners; continue planning for first partners meeting
    - Third quarter: re-write and improve procedures; preparation of demonstration for potential partners
    - Fourth quarter:
        - hold 1 workshop of grant personnel and libraries interested in building archives of digital geospatial data and participating in a national network of such archives; to evaluate work done and to do in coming four quarters (and more generally in second year of grant)
        - followup after workshop, finding ways to improve partner meetings
- Second year
    - First quarter: from lessons learned during the first year, determine potential partners to work with during the second year, and set priorities as to which potential partners to contact first; begin process of getting in touch with new potential partners, and keeping up communication with existing partners; re-work procedures for working with potential and committed partners
    - Second quarter: continue working with potential partners set loading priorities; evaluate procedures for working with potential and committed partners
    - Third quarter: contacting potential partners continues; start writing best-practices documents, aiming toward drafts for partner workshops end of fourth quarter
    - Fourth quarter
        - by end of fourth quarter, have contacted all potential partners on prioritized list developed during first quarter
        - hold a workshop on archiving digital geospatial data, for potential and committed partners interested in participating in a national network of archives of digital geospatial data, with the draft best-practices document as chief topic of work for both workshops

- Third year
    - First quarter: re-write draft best-practices document into a solid Version 1.0 document; make prioritized list of potential partners to be contacted

during third year; develop plans for carrying on the network after the 3-year grant period.
- Second quarter: issue Version 1.0 best-practices document for comment
- Third quarter
  - hold one workshop on archiving digital geospatial data, for potential and committed partners interested in participating in a network of archives of digital geospatial data; Version 1.0 document as chief topic of workshop
- Fourth quarter
  - by end of fourth quarter, have contacted all potential partners on prioritized list developed during first quarter
  - by end of fourth quarter, have "final" version of best-practices document - to include how to become a member of a national network of archives of digital geospatial data - available via the grant's Webpage
  - notify all committed partners of status of the network and of suggested plans (developed during the first quarter) for carrying on and adding to the network after the 3-year grant period is completed

WORK PLAN FOR CONTENT RETENTION/TRANSFER

General Model for Content Retention and Transfer

All content acquired during and under the auspices of the project by the Partners will be ingested by one or another partner, with the intention of storing it permanently. The goal of the Retention phase is to utilize replication to heterogeneous storage repositories. If content can no longer be retained by one partner, the other partner(s) will make every effort to accommodate the content in jeopardy. In the further event that this transfer cannot be accomplished securely, then such content will be transferred to the Library of Congress, as required under the NDIIPP conditions, using transfer protocols and format standards negotiated with the Library of Congress at that time.

The Partners have distinct technologies and programs for preserving digital content as described in the following sections. Over time, it is expected these approaches will benefit, though they will not necessarily converge, from the interaction and cooperation among the Partners and staffs.

Policy and legal issues will be worked out with each collection owner so that there is a clear understanding by the collection owner and by the Partners as to what are each party’s responsibilities and expectations

Economic and technical issues encountered - judging from previous experience in dealing with terabytes of digital geospatial data - will in the main have to do with planning and budgeting for sufficient computer-technical staff and hardware to keep up both with adding new collections and metadata and with maintaining and providing access to existing collections.

UCSB

MIL currently has approximately seven terabytes of digital geospatial data (about 290,000 files) and servers as appropriate, and in early 2004 added an additional five terabytes of storage. All digital geospatial data are regularly backed up on tape and the tapes stored at the library's off-campus storage facility. This pattern of regularly scheduled backup is one that during the contract period will be followed for the archiving of digital geospatial data. In addition, data will be backed up at SDSC. Protection from unauthorized use will be the work of the authentication procedures.

MIL's efforts during the contract are to archive multiple terabytes of different kinds of digital geospatial data and to generate a best-practices document for such archiving, thus encouraging as many institutions as possible to participate in this endeavor. During the grant period, MIL will be archiving many different types of geospatial data, of many different geographic areas, as a part of the work on determining how archiving may best be done. Thus when there are data archived at Santa Barbara during the contract period that may more appropriately in the stewardship of a shareholder other than MIL, MIL will work at turning over the data to appropriate agencies. For example, MIL has a user base for which digital geospatial data of California is of supreme importance and even that data exists in such large amounts that it is common sense to share out the archiving work with other shareholder institutions in California, e.g., government agencies and libraries.

Stanford

Stanford University is strongly committed to the long-term preservation of selected digital information. The current focus for the SDR staff is to save bits and bytes as they arrive in a secure, "enterprise-level" managed environment, assure the integrity of the data received, and rigorous extraction and development of metadata at the point of ingestion. The further vital aspects of digital preservation will be addressed as the state of the art advances and as necessity dictates, once it is assured that the data objects and associated metadata are secure.

Internally, Stanford has been developing a program of service level agreements for locally owned content that would depend on the form of content being delivered, such that "canonical formats" would be guaranteed to be kept readable as well as in bit-perfect replica of the original submission. Given the variety of geospatial data types and formats to be gathered for this project, Stanford assures at least that bit-perfect files will be retained. Where practical and possible, as part of the larger SDR effort, derivatives of files (either forward migrated or "canonicalized") will be available. Stanford will work with the LongNow Foundation through the project to devise, test, and establish remote secure storage of copies of Repository content.

Content Retention and Transfer Milestones

- First year
    - First quarter
        - First 7 weeks:
            - advertise, interview, and hire project staff

            - develop and implement procedures for identifying and selecting data, and for acquiring from content providers data and metadata; discuss with Oddens' Bookmarks, Map History, NASA, ESRI, and other primary sources of digital geospatial data to determine schedule and work flow for obtaining either site URLs or data
            - write form emails/letters to send to owners of digital geospatial data re possible archiving of data
            - Recruit and engage Advisory Group
            - Get Webpage/Webpage architecture for grant set up and live on the Web; write procedures for additions by grant personnel
        - Remaining 6 weeks:
            - first week: orientation for new staff
            - next 5 weeks: train new staff and document processes
    - Second quarter: programming staff from UCSB travel to Stanford to assist in loading ADL-Operational on Stanford servers; pilot tests of obtaining content from content providers (e.g., California state and selected county government agencies; URLs obtained from Oddens' Bookmarks and Map History; loading of data and metadata into UCSB ADL begins; begin writing procedures for content identification, selection, and acquisition (including loading procedures)
    - Third quarter: pilot tests of loading data and metadata into ADL at Stanford (programmers from UCSB to go to Stanford to assist); re-write and improve procedures; prepare demonstration for Advisory Board and workshop of grant personnel
    - Fourth quarter:
        - by end of fourth quarter, aim to have two terabytes of archived digital geospatial data
        - hold 1 workshops of grant personnel and libraries interested in building archives of digital geospatial data and participating in a national network of such archives
        - following workshop, hold meeting of Advisory Board
- Second year
    - First quarter: from lessons learned during the first year, determine content identification, selection, and acquisition priorities for each quarter of second year of grant; identify and select content as per parameters; obtain data and metadata
    - Second quarter: fine-tune identification and selection parameters; set loading priorities; evaluate loading procedures for data and metadata
    - Third quarter: content identification, selection, and acquisition continues; start writing best-practices documents, aiming toward draft documents for workshops held at end of fourth quarter; best-practices document to include how other libraries may load ADL-Operational and build their own geospatial-data archives

    - Fourth quarter
        - by end of fourth quarter, aim to have four terabytes (total) of archived digital geospatial data
        - hold 2 workshops on archiving digital geospatial data, one for grant personnel and one for libraries interested in building digital geospatial-data archives and participating in a national network of archives of digital geospatial data, with the draft best-practices document as chief topic of work for both workshops
        - hold meeting of Advisory Board

- Third year
    - First quarter: re-write draft best-practices document into a solid Version 1.0 document; continue content identification, selection, and acquisitions
    - Second quarter: issue Version 1.0 best-practices document for comment
    - Third quarter
        - hold one workshop on archiving digital geospatial data, for libraries interested in building archives of digital geospatial data and in participating in a network of archives of digital geospatial data; Version 1.0 document as chief topic of workshop
        - hold meeting of Advisory Board
    - Fourth quarter
        - by end of fourth quarter, have six terabytes (total) of archived digital geospatial data
        - by end of fourth quarter, have "final" version of best-practices document *to include how to become a member of a national network of archives of digital geospatial data* available via the grant's Webpage
        - write all final reports.

SUMMARY OF APPROVED DELIVERABLES

Proposed deliverables for the joint project will be:

- A national network of heterogeneous federated geospatial repositories with a goal of at least four nodes.
    - A series of protocols governing partnership of partners of the network, including but not limited to: criteria for content; communication obligations of partners; transfer of data to the Library, upon the Library's request; etc.
- Archived digital content of approximately 12 to 15 terabytes of geospatial data with associated metadata capable of being transferred to or replicated at the Library of Congress

- Preparation of Best Practices papers, including addressing metadata for geospatial files, work flows for ingestion of geospatial files, and interoperation among repositories
- A model partnership agreement for digital – particularly geospatial – distributed archives
- A *gap analysis* report on future needs for geospatial digital archiving
- A collaborative website for the National Geospatial Federated Digital Repository that provides archives, links, best-practices documents, FAQs, and contacts information
- Any software developed by UCSB or Stanford in the performance of the grant, including but not limited to middleware, metadata ingest or archives management.